Dealing with Randomness and Concept Drift in Large Datasets

MWITONDI, Kassim S. <http://orcid.org/0000-0003-1134-547X> and SAID, Raed A.

data

MDPI

Article

Dealing with Randomness and Concept Drift in Large Datasets

Kassim S. Mwitondi [1,*] and Raed A. Said [2]

[1] College of Business, Technology & Engineering, Sheffield Hallam University, Industry & Innovation Research Institute, 9410 Cantor Building, City Campus, 153 Arundel Street, Sheffield S1 2NU, UK
[2] Faculty of Management, Canadian University Dubai, Al Safa Street-Al Wasl, City Walk Mall, Dubai P.O. Box 415053, United Arab Emirates; raed.saeed@cud.ac.ae
* Correspondence: k.mwitondi@shu.ac.uk

Abstract: Data-driven solutions to societal challenges continue to bring new dimensions to our daily lives. For example, while good-quality education is a well-acknowledged foundation of sustainable development, innovation and creativity, variations in student attainment and general performance remain commonplace. Developing data -driven solutions hinges on two fronts-technical and application. The former relates to the modelling perspective, where two of the major challenges are the impact of data randomness and general variations in definitions, typically referred to as concept drift in machine learning. The latter relates to devising data-driven solutions to address real-life challenges such as identifying potential triggers of pedagogical performance, which aligns with the Sustainable Development Goal (SDG) #4-Quality Education. A total of 3145 pedagogical data points were obtained from the central data collection platform for the United Arab Emirates (UAE) Ministry of Education (MoE). Using simple data visualisation and machine learning techniques via a generic algorithm for sampling, measuring and assessing, the paper highlights research pathways for educationists and data scientists to attain unified goals in an interdisciplinary context. Its novelty derives from embedded capacity to address data randomness and concept drift by minimising modelling variations and yielding consistent results across samples. Results show that intricate relationships among data attributes describe the invariant conditions that practitioners in the two overlapping fields of data science and education must identify.

Keywords: artificial neural networks (ANNs); Big Data; concept drift; data science; supervised modelling; sustainable development goals; unsupervised modelling

Citation: Mwitondi, K.S.; Said, R.A. Dealing with Randomness and Concept Drift in Large Datasets. *Data* **2021**, *6*, 77. https://doi.org/10.3390/data6070077

Academic Editors: Donatella Merlini, Maria Cecilia Verri, Leonardo Grilli and Carla Rampichini

Received: 21 May 2021
Accepted: 9 July 2021
Published: 19 July 2021

1. Introduction

Many studies have been carried out to try and establish the right breaking point for university pedagogy innovations, using a combination of tools, techniques and skills [1]. The data deluge is producing more challenges and opportunities in these areas, as the sectoral data generation typically outpaces processing capacities. In higher education, research has focused on identifying the triggers of pedagogical performance [2] while in the machine learning field, studies have widely varied-from observational [3,4] to utilising limited information to predict performance [5]. In both cases, the basic hypotheses have been to seek potential solutions from multiple data attributes-some undiscernible, hidden in prior and posterial learning environment, outcome and purpose, which can be uncovered through a combined technical and soft interdisciplinary approach. Migueis et al. [5] proposed a two-stage model, based on information at the end of the students' first academic year to predict their overall academic performance. The downside in their approach was that the study was confined to a single Engineering School and also that using information at the end of the year as a single predictor could not be justified for robustness. A number of interdisciplinary studies involving pedagogy and machine learning have addressed variations associated with datasets, human intervention and artificial intelligence algorithms [6]. Such variations arise from data sampling and they can have a negative impact on predictive

modelling results as they tend to degrade model performance, if static relationships are assumed [7].

We propose a novel approach to extracting pedagogical performance data from a cross section of university students, via addressing data randomness and concept drift [8]-changes in underlying data relationships over time. Unlike in Brooks et al. [4] and Hua Leong and Marshall [3], who applied machine learning techniques to monitor students' engagement in class, we shall be taking a more robust approach to searching for patterns in historical data. We apply machine learning techniques with data from the United Arab Emirates (UAE) and we demonstrate how discovering such information helps educationists, researchers and other stakeholders to forge unified efforts in uncovering and analysing relevant data in an interdisciplinary context. Pedagogical data were obtained from CHEDS [9]-a central data collection platform for the United Arab Emirates (UAE) Ministry of Education (MoE) in a standard Relational Database format used for storing and accessing very large datasets. A crucial aspect of this work relates to the intricate relationships among data attributes that practitioners in the two overlapping fields of data science and education must identify. In particular, it presents two scenarios of the same target variable, based on slightly different definitions and hence highlighting the impact of concept drift [10,11], a well-documented issue in machine learning.

The application aligns with the path for the UAE's Commission for Academic Accreditation (CAA), a body within the Ministry of Education (MoE), which is responsible for licensing academic institutions across the Emirates. It is responsible for study programs and awarding academic qualifications in accordance with the Standards for Institutional Licensure and Program Accreditation (SILPA) [12]. Among its stipulations, SILPA requires that workload assignments be equitable and reasonable and that they include the entire range of a faculty member's responsibilities, such as instruction, advising, project supervision, internship supervision, independent study, committee work, thesis or dissertation supervision, guidance of student organisations, research, service, and curriculum development. Through career service centres, students are supported in multitasking their learning experiences as well as their internship while ensuring that a good relationship is maintained between the industry and the community.

This study is planned and executed in accordance with the rules laid down by the CAA and it is organised as follows. Section 1 presents the study background, motivation, objectives, gap challenges and a brief review of the relevant literature. Section 2 details the proposed approach, data sources, randomness and concept drift. It also features the learning algorithm used in this study, with illustrations of unsupervised and supervised modelling. Section 3 presents analyses, results and evaluation based on data visualisation, unsupervised and supervised modelling. Section 4 outlines the paper's novelty and discussions and finally, Section 5 presents concluding remarks.

1.1. Motivation

Recent developments in data generation and processing have ignited new dimensions in tackling the challenges we face, which, in the education sector, typically hinge on the variations in attainment and performance among learners. Data attributes on learning and assessment dynamically interact with other multi-faceted data attributes that are often hidden in the learners' prior and posterior learning environment, outcome and purpose. CHEDS [9], a data store that provides a central data collection platform for the MoE, requires universities that were accredited by the CAA [12] to constantly submit their data to them. In particular, it requires that internship data be collected at the end of each fall semester, using an especially designed template providing information about each intern during the academic year.

1.2. Underlying Problem, Aim and Objectives

The aim and objectives of this work were inspired by a technical data science perspective on the one hand, and educational performance and its related impact on society

on the other. More specifically, it seeks to uncover an interdisciplinary approach to addressing issues of data randomness [7,13] and concept drift [11] in education, by exploring different scenarios of using multiple data samples from the same source. Data generation, acquisition, storage, analytics, visualisation and utilisation for performance enhancement appear as key attributes of studies in various sectors [14–16]. Consequently, the work will be highlighting paths towards a meaningful application of data analytics techniques-from simple data visualisation (DV) to complex machine learning applications—via an integrated generic algorithm and educational performance data attributes. Its key objectives are as follows:

1 To address data randomness and concept drift in a real-life application;
2 To apply the sample–measure–assess (SMA) algorithm for unsupervised and supervised model optimisation;
3 To highlight pathways for educationists, data scientists and other researchers to follow in engaging policy makers, development stakeholders and the general public in putting generated data to use.
4 To motivate an unified and interdisciplinary understanding of data-driven decisions across disciplines.

1.3. Gap Challenges

Application of tools and techniques in addressing societal challenges has always been part and parcel of human development. Enhancement in the ways we generate and consume data provide both challenges and opportunities for researchers, constantly filling and opening knowledge gaps. The paper seeks to address knowledge gaps in both the broad area of managing higher education through data-driven solutions and in predictive modelling. Modelling pedagogical performance feeds into Sustainable Development Goal (SDG) #4-Quality Education-while addressing data randomness and concept drift hinges on model optimisation-a classical challenge in making data-driven decisions [7]. In the next section, we outline the proposed modelling approach that seeks to bridge these challenges.

2. Proposed Approach

The efficacy of data-driven decision making is conditional on a wide range of factors. Typically, data, tools, techniques and skills combine to provide the ability to research and identify what is working well and what is not. This section describes the approach adopted in this study-a coherent combination of the foregoing factors. Most importantly, it highlights the underlying variations in the general behaviour of both data and methods, as used in real-life applications.

2.1. Data Sources

Table 1 presents the data attributes used in this work. The original data, consisting of 3145 observations on 19 variables, came from a university data repository [9] in the UAE over an 11-year period (2005–2016). Students were from different countries across the Middle East, with the majority being from the United Arab Emirates (56%) and Oman (28%). Other nationalities included Palestine (3.7%), Jordan (3%), Syria (3.2%), Yemen (1.5%), Egypt (1%), Iraq (1%) and a few from Lebanon, Somalia, Qatar and the UK. About 99.2% were from the Middle East.

Business administration, education, engineering and information technology, law, pharmacy and early years teaching, constituted 19.8%, 24.9%, 1.1%, 45.8%, 4.4% and 4.0% of the courses undertaken by the students, respectively. Note that while the data cover a period long enough to represent reliable and stable patterns, this does not rule out significant data dynamics in the future-which are very likely under current Big Data conditions. The data attributes in Table 1 are static, linked to a known population under study-in this case, the university superset in the UAE, in which the overall behaviour of pedagogical performance is presumed to be known (hypothesised). This population will typically generate and consume a lot of data which can be used to test the dynamics of

their overall behaviour over time. We highlighted the potential new research directions in Section 5 to be guided by such dynamics. For the purpose of supervised modelling, the variable CLS in Table 1 was added as a discretisation of GPA performance, giving two instances—binary and trinomial. This discretisation results in class labels and its general implication is that it becomes informative of the natural groupings in the data.

Table 1. Selected students' data attributes.

Code	Variable	Type	Description	Summaries
IST	Institution	String	University	One with two campuses
GDR	Sex	Binary	Sex	Female (55%); Male (45%)
NTA	Nationality	String	Home country	UAE (56%) Oman (28%)
CPS TYP	Type	String	Start or cont/trans	Bach (74%); Dip (25.7%); Master's (0.3%)
LVL	Level	String	Diploma, first or post	3 different levels
SPC	Specialisation	String	Broad specialisation	5 different specialisations
MJR	Major	String	Specific field	43 different major subjects
INT	InternSector	String	Internship sector	60 different sectors
PCD	ProgramCredits	Numeric	Total credits to grad.	Q1 = 24 Med = 129 Mean = 102 Q3 = 129
RCP	RegCreditsPrev	Numeric	Reg. Spring credits	Q1 = 12 Med = 15 Mean = 14 Q3 = 16
PVC	PrevCreditsComplete	Numeric	Comp. spring credits	Q1 = 12 Med = 15 Mean = 13.1 Q3 = 15
RGC	RegCredits	Numeric	Reg. Curr. credits	Q1 = 9 Med = 15 Mean = 12.6 Q3 = 16
CMC	CumulativeCredits	Numeric	Cumulative credits	Q1 = 15 Med = 93 Mean = 76 Q3 = 108
CGP	CumulativeGPA	Numeric	Cumulative GPA	Q1 = 2.2 Med = 2.6 Mean = 2.7 Q3 = 3.1
QES	QualifyingExitScore	Percentage	Score from Q-Award	Q1 = 65 Med = 74 Mean = 68 Q3 = 82
BSG	BeforeSemGPA	Numeric	GPA Before internship	Q1 = 2.2 Med = 2.8 Mean = 2.7 Q3 = 3.4
ISG	InSemGPA	Numeric	In-semester GPA	Q1 = 2.7 Med = 3.1 Mean = 3.1 Q3 = 3.6
ASG	AfterSemGPA	Numeric	GPA After internship	Q1 = 2.3 Med = 3.0 Mean = 2.8 Q3 = 3.5
CLS	Class	Binary	Tweaked GPA	$\geq$Mean (49%) and $<$Mean (51%)

2.2. Data Randomness and Concept Drift

Table 2 exhibits allocation rule errors due to data randomness. The first column represents the presumed population error which a trained model seeks to reproduce by learning a rule from the training data (column 2), validating it on validation data (column 3) and testing it on newly previously unseen data (column 4). Randomness in training, validation and test data is quite crucial while dealing with large datasets, since it makes model optimisation a natural challenge to data modelling [7,13]. The impact of the variable CLS in Table 1 on the decision-making process depends on the way it was done, the reasoning behind it and its potential consequences. It reflects concept drift in the sense that discretisation criteria may change over time, and as such, it must be accomplished with a thorough and comprehensive anticipation of its consequences. While its definition will typically be guided by expert knowledge, variations may arise due to different circumstances, rendering it inherently random.

Table 2. Allocation rule errors due to data randomness.

Population Error	Training Error	Cross Validation Error	Test Error
$\psi_{D,POP}$	$\psi_{D,TRN}$	$\psi_{D,XVD}$	$\psi_{D,TEST}$
Actual population error	From random training	From random validation	From random testing

The process of converting data to knowledge naturally derives from standard statistical sampling, modelling and evaluation [17,18]. This paper shall be applying multiple models, based on Algorithm 1 [19] to address the randomness in Table 2. In line with objective #2 in Section 1.2, the algorithm was designed to lead to a generalised strategy aimed at striking a balance between model accuracy and reliability across samples and applications.

2.3. Learning Rules from Data by Sampling, Measuring and Assessing

This section describes Algorithm 1, the mechanics of which were previously published in related work [19–21]. The algorithm was designed to address data randomness based on built mechanics for minimising variation and its potential to yield consistent results across samples. Without loss of generality, the algorithm assumes a relational database model in which the data source $\mathbf{X} = [x_{i,j}]$ is organised in rows and columns. The algorithm operates on the designated accessible dataset $\mathbf{X} = [x_{i,j}]$ in Table 1, applying any relevant learning model, typically defined as

$$F(\phi) = \underbrace{P}_{x,y\sim\mathscr{D}}[\phi(x) \neq y] \tag{1}$$

where $\mathscr{D}$ is the underlying distribution and $P[\phi(x) \neq y]$ denotes the probability of a predicted value not being equal to the true value. The model in Equation (1) describes a supervised case scenario, but it can also take an unsupervised form by removing the class label and focusing on the similarity and/or dissimilarity of the data points in $\mathbf{X} = [x_{i,j}]$.

In both cases, allocation rules are learnt through Algorithm 1 below. Initialisation of s as a percentage of $[x_{v,\tau}]$, and the choice of K is application specific, to be decided by the investigator. The constant κ used here is a free parameter, determined by the user. The main idea of the algorithm is to generate random samples for model training, validation and testing purposes, varying key parameters from sample to sample as outlined below.

Algorithm 1 SMA—Sample, Measure, Assess

1: **procedure** SMA
2: Set $\mathbf{X} = [x_{i,j}]$: Accessible Data Source
3: Learn $F(\phi) = \underbrace{P}_{x,y\sim\mathscr{D}}[\phi(x) \neq y]$ based on a chosen learning model
4: Set the number of iterations to a large number K
5: **Initialise:** $\Theta_{tr} := \Theta_{tr}(.)$: Training Parameters
6: **Initialise:** $\Theta_{ts} := \Theta_{ts}(.)$: Testing Parameters
7: **Initialise:** $\Pi_{cp} := \Pi_{cp}(.)$: Comparative Parameters
8: **Initialise:** s as a percentage of $[x_{v,\tau}]$, say 1%
9: s_{tr} : Training Sample $[x_{v,\tau}] \leftarrow [x_{i,j}]$ extracted from $\mathbf{X} = [x_{i,j}]$
10: s_{ts} : Test Sample $[x_{\bar{v},\tau}] \leftarrow \left[x_{l\neq i,j}\right]$ extracted from $\mathbf{X} = [x_{i,j}]$
11: **for** $i := 1 \rightarrow K$ **do**: Set K large and iterate in search of optimal values
12: **while** $s \leq 50\%$ of $[x_{v,\tau}]$ **do** Vary sample sizes to up to the nearest integer 50% of X
13: **Sampling for Training:** $s_{tr} \leftarrow X$
14: **Sampling for Testing:** $s_{ts} \leftarrow X$
15: **Fit Training and Testing Models** $\hat{\mathcal{L}}_{tr,ts} \propto \Phi(.)_{tr,ts}$ with current parameters
16: **Update Training Parameters:** $\Theta_{tr}(.) \leftarrow \Theta_{tr}$
17: **Update Testing Parameters:** $\Theta_{ts}(.) \leftarrow \Theta_{ts}$
18: **Compare:** $\Phi(.)_{tr}$ with $\Phi(.)_{ts}$: Plotting or otherwise
19: **Update Comparative Parameters:** $\Pi(.)_{cp} \leftarrow \Phi(.)_{tr,ts}$
20: **Assess:** $P(\Psi_{D,POP} \geq \Psi_{D,TRN}) = 1 \Longleftrightarrow \mathbb{E}[\Psi_{D,POP} - \Psi_{D,TRN}] = \mathbb{E}[\Delta] \geq 0$
21: **end while**
22: **end for**
23: **Output the Best Models** $\hat{\mathcal{L}}_{tr,ts}$ based on $\mathbb{E}[\Delta] \geq 0$
24: **end procedure**

The notations $\Theta_{tr} := \Theta_{tr}(.)$ in step 5 and $\Theta_{ts} := \Theta_{ts}(.)$ in step 6 are model-specific and they represent the initialisation of model-specific parameters. For example, initial training parameters for a finite mixture model of two normals with different variances would be $\Theta_{tr} = \{\mu_1, \mu_2; \sigma_1, \sigma_2\}$. For an artificial neural network model, it may include model architecture-related parameters such as the number of layers, neurons, learning rate, etc., while for a decision tree, it could be the tree depth, splitting criterion, measure

of purity, etc. In the general context of predictive modelling, Θ_{tr} would be assigned prior probabilities of class membership and respective distributional densities of the classes. More specifically, $\Theta_{tr} := \Theta_{tr}(.)$ denotes the assignment of initial training parameters to Θ_{tr} and $\Theta_{tr}(.)$ denotes the source of such parameters-as a function of expert domain knowledge, prior information or exploratory data analysis.

While Algorithm 1 clearly follows the standard machine learning pipelining, i.e., data processing, learning, evaluation and prediction [22], novelty lies in its ability to address data randomness through its mechanics. More specifically, the updating of the training and testing parameters $[\Theta_{tr}(.) \leftarrow \Theta_{tr}$ and $\Theta_{ts}(.) \leftarrow \Theta_{ts}]$ occur alongside randomly drawn samples, $[x_{\nu,\tau}] \leftarrow [x_{i,j}]$ and $[x_{\nu,\tau}] \leftarrow \left[x_{l\neq i,j}\right]$ at steps 8 through 10. The samples are random and remain stateless across all iterations. Multiple machine learning models $\hat{\mathcal{L}}_{tr,ts} \propto \Phi(.)_{tr,ts}$ are fitted, compared and updated over iterations 11–19. At step 20, the best performing model was selected based on the comparison between the probability of the population error and that of the training error $[P(\Psi_{D,POP} \geq \Psi_{D,TRN})]$.

In addressing issues of data randomness, Algorithm 1 draws from existing modelling techniques such as the standard variants of cross-validation [23] and permutation feature importance [24]. It is also comparable to various models of bagging and bootstrapping. It is known that the aggregation of classifiers based on a generic bagging procedure may not always lead to the best solutions while bootstrapping without an underlying model but only relying on sample representativeness which may not always be guaranteed [25]. Unlike these methods, Algorithm 1 has a built–in mechanism that allows it to handle data randomness more efficiently, as evidenced by its previous applications [19,25].

2.4. Experimental Setup

Exactly how we deal with the data attributes in Table 1 will depend on the data types and our initial position of what we are looking for in the data. This understanding is guided by the problem space and study objectives-the general rule of thumb is that it requires a strong understanding of the system being investigated. This section summarises the experimental setup based on three scenarios-data visualisation, unsupervised and supervised learning.

2.4.1. Data Visualisation

Data visualisation is a prolific way for gaining insights into the distributional behaviour of the data used in any analysis. Visually conveying patterns of interest from our data leverages analyses and interpretations that follow. Data visualisation is becoming increasingly popular in the modern era of Big Data, as a tool for gaining insights into large volumes of data we generate every day. Visualising patterns and trends helps internalise the situation, curating data into a form easier to understand and highlighting blind spots for analyses. Thus, we will be visualising the data in Table 1 as the first step in order to gain a good understanding of its behaviour. The objectives in Section 1.2, the gap challenges in Section 1.3 and the visual patterns from the data in Table 1 help to determine the level of detail and validity of results from unsupervised and supervised modelling, the mechanics of which are outlined below.

2.4.2. Unsupervised Modelling

We use principal component analysis (PCA), a dimensional reduction technique, to illustrate how Algorithm 1 can learn rules from unlabelled data. We used ten numeric variables in Table 1 and PCA to repeatedly sample GPA and students' credits data to reduce the data dimensionality and be evaluated on any of the other attributes as follows:

$$\mathcal{I} = \{\text{PCD, RCP, PVC, RGC, CMC, CGP, QES, BSG, ISG, ASG}\} \subset \mathbb{R}^n \tag{2}$$

Given the ten variables in Equation (2) as numerical performance indicators, we can apply PCA to extract a maximum of 10 random components, irrespective of the sample

sizes. In this sense, each extracted component is therefore a random linear combination of all variables-an estimated weighted sum such that:

$$\mathcal{PC}_k = \{w_k\text{PCD}, w_k\text{RCP}, w_k\text{PVC}, w_k\text{RGC}, w_k\text{CMC}, w_k\text{CGP}, w_k\text{QES}, w_k\text{BSG}, w_k\text{ISG}, w_k\text{ASG}\} \quad (3)$$

where $k = 1, 2, 3, ..., 10$. The vectors w_{ik} are chosen such that the following conditions are met:

1. Each of the determinants equals 1, $\|w_k\| = 1$;
2. Each of the $\mathcal{PC}_k$, maximises the variance $V\left\{w_k'\mathcal{I}_k\right\}$; and
3. The covariance $COV\left\{w_k'\mathcal{I}_k\, w_r{}'\mathcal{I}_r\right\} = 0$, $\forall k < r$.

In this section, PCA is applied in establishing the underlying components in the data, created through linear combinations of the variables in Table 1. The application is designed to provide insights into naturally arising structures in data alongside other DV results, as a precursor to predictive modelling, without running through the SMA algorithm.

2.4.3. Supervised Modelling

As mentioned above, a discretised variable, CLS, was added to the data in Table 1, giving a discretised vector of GPA performance, as binary and trinomial. The two class labels were created based on the rule in Equation (4):

$$\text{CLS} = \begin{cases} \text{High:} & \text{If} \quad \sum_i^N V_i(\text{BSG, ISG, ASG} >) \,/\, \text{Length}V_i \geq \text{Mean}V_i \\ \text{Low:} & \text{Otherwise} \end{cases} \quad (4)$$

Equation (4) creates a binary case of class labels, yielding High if each of the data entries in the vector variable is more or less than the vector average and Low otherwise. This discretisation can be extended to any multinomial case and variable, depending on the problem at hand. In this case, the decision was based on the rationale gained from the findings in Sections 3.1 and 3.2. The trinomial instance, in Equation (5), followed a similar rule, except that rather than using the mean as the cut-off point, quartiles were used:

$$\text{CLS} = \begin{cases} \text{High:} & \text{If} \quad \sum_i^N V_i(\text{BSG, ISG, ASG} >) \,/\, \text{Length}V_i \geq \text{Quartile}_3 V_i \\ \text{Medium:} & \text{If} \quad \text{Quartile}_1 V_i \leq \sum_i^N V_i(\text{BSG, ISG, ASG} >) \,/\, \text{Length}V_i < \text{Quartile}_3 V_i \\ \text{Low:} & \text{If} \quad \sum_i^N V_i(\text{BSG, ISG, ASG} >) \,/\, \text{Length}V_i < \text{Quartile}_1 V_i \end{cases} \quad (5)$$

The choice of the variable and form of discretisation will depend on the underlying problem, and most importantly, on the initial insights gained from either expert knowledge, exploratory data analysis (EDA), data visualisation (DV), unsupervised modelling or a combination of each. For supervised analyses, the simplest rule would be to average the GPA-related variables BSG, ISG and ASG to create a binary class variable, describing each student's GPA attainment as above average (A) and below average (B) or a multinomial class. The 32 different levels of INT can be discretised by combining sectors-e.g., technology, business and law enforcement. For both unsupervised and supervised modelling purposes, the levels for this class variable can be varied as a yardstick for measuring associations among the 19 data attributes, hence providing insights into the tuning parameters that education policy makers need to focus on.

Now, if we set $CLS \propto F(\phi)$ in SMA Algorithm 1, we can adopt the Bayesian approach to use existing prior knowledge to learn more about the data and generate new (posterior) knowledge-i.e, we are interested in the posterior probability of a particular event, e.g., belonging to the kth class given evidence in the data as in Equation (6):

$$f(\text{CLS} = k|x) = \frac{f(\text{CLS}|x)\pi_k}{\int_{-\infty}^{\infty} f(\text{CLS}|x)\pi_k \text{dCLS}} \propto \frac{f_k(x)\pi_k}{\sum_{k=1}^K f_k(x)\pi_k} \propto P(\text{CLS}|x) = \frac{P(x|\text{CLS})P(\text{CLS})}{P(x)} \quad (6)$$

where $f(x)$ and π_k are the data density and class priors, proportional to $P(x|\text{CLS})$ and $P(\text{CLS})$, respectively, both of which are estimated from data and thus inevitably generating

the errors in Table 2. Equation (6) represents the Bayesian rule which we can use to define the overall misclassification error for each one of the errors in Table 2 as the sum of the weighted probabilities of observing data belonging to a particular class given that we are not in that class-meaning that:

$$\Psi_{D,XVD} = \sum_{j}^{k} \sum_{i}^{n} \pi_j P(x_i \in \text{CLS}_j | \text{CLS}_i) \tag{7}$$

Any adopted machine learning model can be optimised by harmonising data variability through cross-validation using the SMA Algorithm [20,21], which learns a model $F(\phi) = \underbrace{P}_{x,CLS\sim\mathscr{D}} [\phi(x) \neq \text{CLS}]$, where $\mathscr{D}$ is the underlying distribution, and it provides the mechanics for assessing the models. Given labelled data, its outputs provide great insights into the overall behaviour of the data, particularly how the attributes relate to the target variable.

In Section 3.3, we apply artificial neural network (ANN) [26,27] deep learning to the data-i.e., training and testing on random samples $s_{tr} = [x_{\nu,\tau}] \leftarrow [x_{i,j}]$ and $s_{ts} = [x_{\nu,\tau}] \leftarrow \left[x_{l\neq i,j}\right]$, respectively. The simplest form of the model is:

$$F(\phi) = w_1 x_1 + w_2 x_2 + ... + w_n x_n + w_n x_{n+1} = \sum_{i=1}^{n} w_i x_i = \lambda \tag{8}$$

where λ is a constant, w_i are weights on the feature vector x_i, and each feature vector satisfies one of the three conditions: $F(\phi) > \lambda$, $F(\phi) = \lambda$ and $F(\phi) < \lambda$. The summation in Equation (8) combines weights and feature vectors to produce an outcome, which describes the standard linear discriminant network. If we denote the corresponding threshold by t_1, then we can re-write the weighted sum in Equation (8) as $f(\phi) = \sum_{i=0}^{n^*} w_{1i} x_i$ where $f(\phi) = u_1 - t_1$ and the output $F(\phi) = f(\theta_1 - t_1)$, with added new synaptic link with $x_0 = -1$ and $w_{10} = t_1$. However, real-life applications are rarely linear in nature and hence, the examples in Section 3.3 are non-linear, typically tracking the linear gradient descent in search of local minima by updating the weights w_i at each step.

The transition from linearity to non-linearity is achieved via activation functions—one of the most popular being the sigmoid, in its logistic form $f(\phi) = \frac{1}{1+\exp(-\alpha x)}$. Varying the slope parameter α yields sigmoid functions of different slopes, continuous and hence differentiable in all points over the interval $[0, 1]$. As the weights (w_i) increase, there is a gradual drift from a linear to a non-linear model, i.e., via the function, ANN have an embedded mechanism for coping with linear, near-linear and non-linear problems—a statistically appealing property. Thus, a univariate quantitative approximation of CLS can be approximated from the logarithmic sigmoidal function in Equation (9):

$$\mathcal{N}_n(x) = \sum_{j=0}^{n} c_j \sigma\left[\langle \alpha_j \,.\, x \rangle + \beta_j\right],\ x \in \mathbb{R}^s\ s \in \mathbb{N} \tag{9}$$

where for $0 \leq j < n$, $\beta_j \in \mathbb{R}$ are the thresholds, $\alpha_j \in \mathbb{R}^s$ are the connection weights, $c_j \in \mathbb{R}$ are the coefficients, $< \alpha_j \,.\, x >$ is the inner product of α_j and x and σ is the activation function of the network. A major challenge in ANN applications is determining the architecture, i.e., the number of hidden layers and the number of neurons. Prediction accuracy changes profoundly with changes in these parameters. It is in this context that the SMA Algorithm is relevant. Through the algorithm, we seek to approximate this function to a high degree of accuracy and reliability, by considering the general case of n, η and ρ neurons in the input, hidden and output layers, respectively. The weights $w_{ik}\ k = 1, 2, \ldots, \eta$ link the input to the hidden layer nodes, while $z_{kj}\ k = 1, 2, \ldots, \rho$ link the hidden layer nodes to the output layer nodes. If we denote the hidden layer output by $\eta_k = f(x, w_k)$, we can define the final output as

$$\text{CLS} \propto F(\phi) = g\left(\sum_{k=1}^{\eta} \eta_k z_{ki}\right) = g\left[\sum_{k=1}^{\eta} z_{ki} f\left(\sum_{i=1}^{n} x_i w_{ik}\right)\right] \tag{10}$$

In the next exposition, we carry out a sequence of analyses from EDA through unsupervised to supervised modelling.

3. Analyses, Results and Evaluation

For the analyses, we follow Reyes [15] who points to the hierarchy of information flow from students to other stakeholders who are capable of providing inputs into the enhancement of the learning processes. We also adapt the Big Data modelling of Sustainable Development Goals (BDMSDG) [20,21], in a down-scaled context for modelling large datasets. Both pathways home in towards efficient decision making by highlighting the potentially useful information in the 19 attributes that stakeholders can meaningfully utilise, hence jointly fulfilling the objectives in Section 1.2.

3.1. Data Visualisation

The top two panels in Figure 1 exhibit the total number of credits on students' transcript counting towards graduation (PCD), with the density plot, on the right hand side panel, showing three or four clear groupings. The bottom panels represent the qualifying exit score (QES), i.e., the students' pre-university GPA. Students captured within these structures are associated with each of the attributes in Table 1, each with its own structure, forming a complex web of cross-attribute relationships. Notice that these patterns are likely to vary widely across samples. Using machine learning techniques, educators and potential employers can uncover more fine-grained patterns and make timely informative decisions.

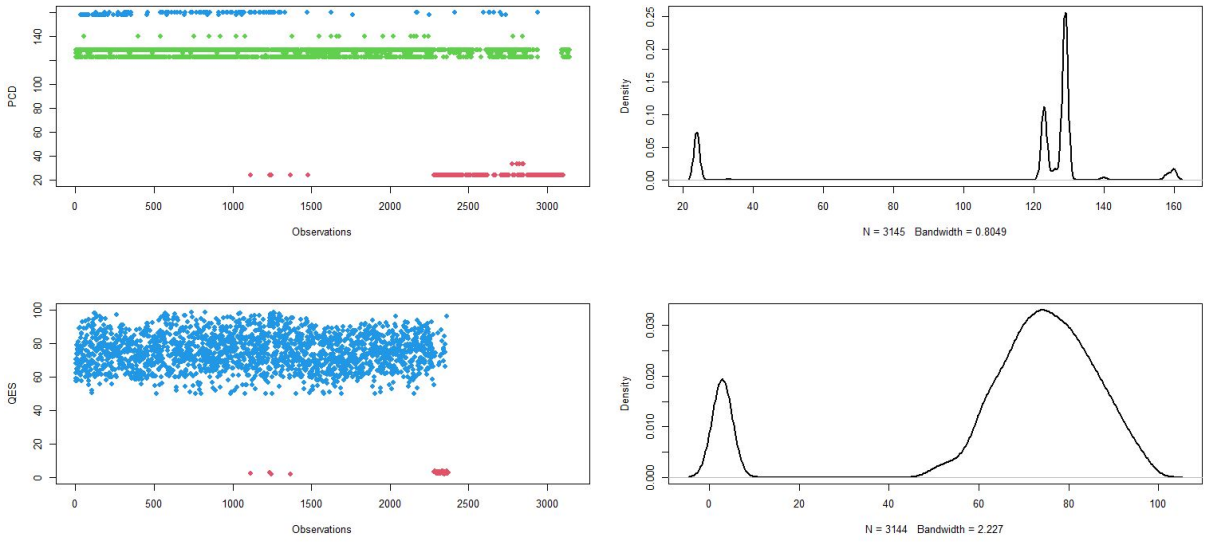

Figure 1. Total credits (PCD) (**top panels**) and the qualifying exit score (QES) (**bottom panels**).

Visualising data in this way provides useful insights for modellers in deciding upon the choice of modelling techniques and ultimately interpretations. At a more granular level, we can examine the individual states of each student based on these patterns alongside other data points. However, due to the sampling-dependent randomness in Table 2, attaining such goals presents both challenges and opportunities to the stakeholders, from both technical and non-technical perspectives, hence the need for adopting interdisciplinary approaches to analyses [16].

In addition to randomness, data overlap provides another challenge. Many real-life applications present far more overlapping cases than these, hence constituting a modelling challenge, the outcomes of which inevitably require a combined modelling and domain knowledge to comprehend. Figure 2 illustrates cases of data swamping and masking,

widely studied and well-documented challenges in statistical outlier detection [28,29]. The top two panels in Figure 2 correspond to the credits registered for during the current academic period (RGC), which are clearly multi-modal. The bottom two panels correspond to the cumulative credits over semesters (CMC) and while they exhibit a bi-modal behaviour, there are many cases which may or may not belong to either mode. In both unsupervised and supervised modelling, the separation of similar from dissimilar cases is central to the performance of the adopted models.

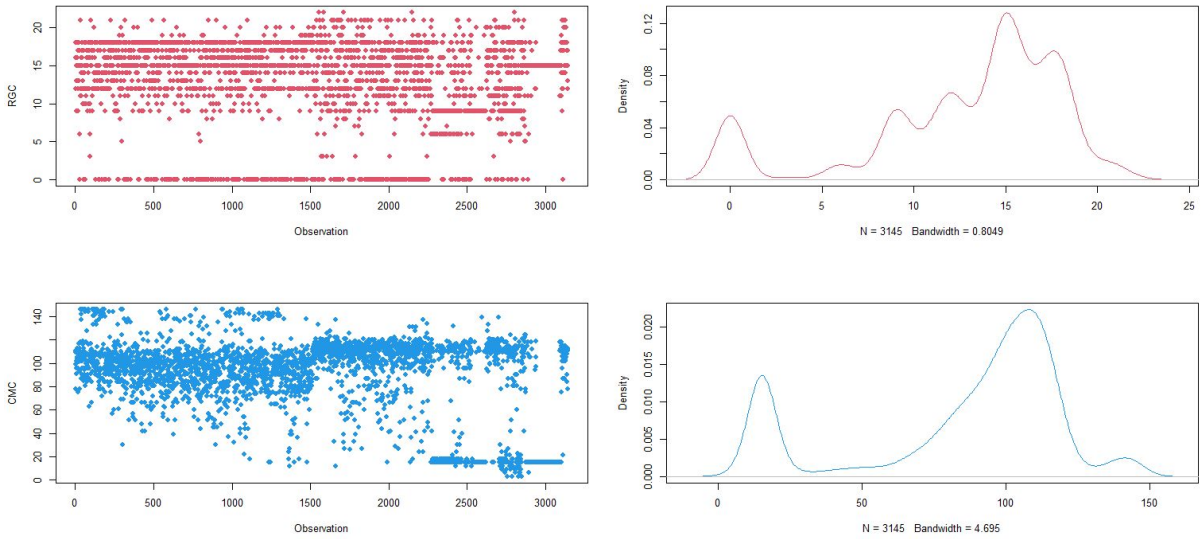

Figure 2. Multi-modal registered credits (RGC) (**top**) and bi-modal cumulative credits (CMC) (**bottom**).

The other important metrics of the data in Table 1 are the GPAs. Figure 3 presents the cumulative GPA from the beginning to the latest enrolment (CGP) as well as the GPAs recorded before the internship, in-semester and after internship, i.e., BSG, ISG and ASG, respectively. The four plots show that there is only a marginal variation among the individually recorded GPAs. Figures 1–3 indicate the typical modelling challenges that real data can present, highlighting the impact of the overall data behaviour on the performance of machine learning techniques. For example, while cases in Figure 1 are well separated, those in Figure 2 are not. Various techniques and approaches are currently in use to iron out the fuzziness, but their adaptation across applications has remained a major challenge. The choice of modelling methods and the interpretation of results are both functions of the randomness in Table 2. In the next sub-section, we apply some of the machine learning methods via the SMA algorithm [20,21].

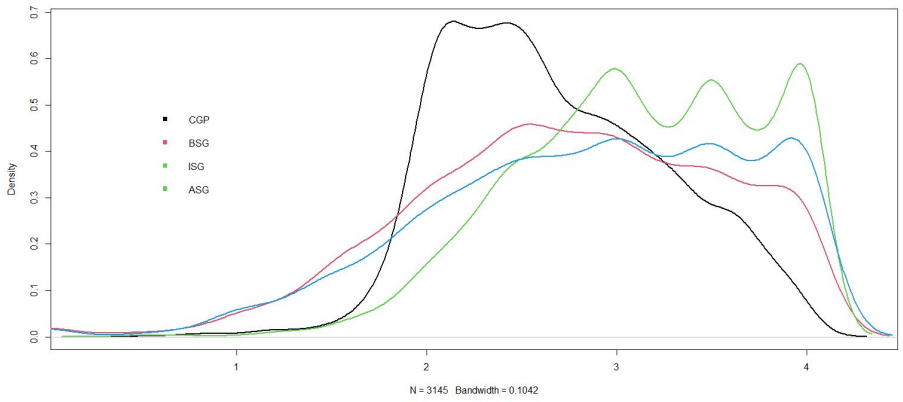

Figure 3. GPA density plots for cumulative, before semester, in-semester and post-internship GPAs.

3.2. Unsupervised Modelling

Equation (3) in Section 2.4.2 generates a total of ten principal components, i.e., natural groupings of the ten numeric variables in Equation (2). The four panels in Figure 4 exhibit bi-plots of the first and second components with respect to GDR, LVL, CLS and SPC with the first three components having eigenvalues of 2.68 1.69 1.53, respectively.

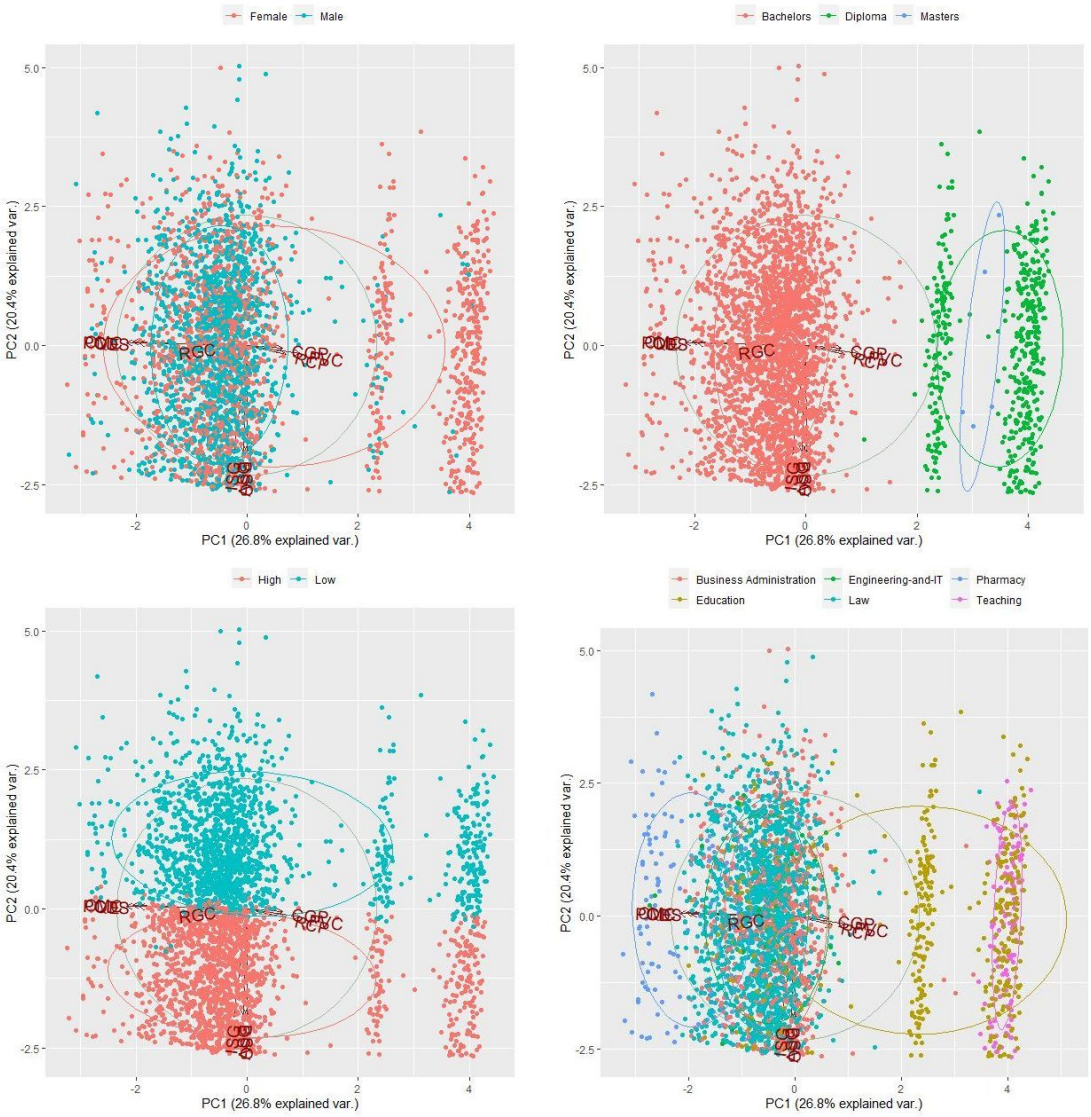

Figure 4. Patterns for the first and second PCA based on sex for the numeric variables in Table 1.

The contribution of each variable in the formation of these groupings is captured via the components' loadings which effectively reduces the dimension of the data in Table 1, providing insights into the variation accounted for by combined variables. Figure 4 indicates three directions for the variables—each with an arrow pointing towards its increasing values. The variables CGP, RCP and PVC point towards high values of the first component, implying that the higher the value of the component, the higher the variables. The three variables BSG, ISG and ASG point towards mid-values of the first component and all the remaining arrows point towards lower values of the component—implying that the lower the value of the component, the higher the variables. Thus, the value of the first component is high if the variables CGP, RCP and PVC are highly scored and it is low if the other variables (except RGC) are highly scored. These directions are reflective of the component loadings (eigenvectors) given in Table 3.

Table 3. PCA rotations for the ten numeric columns in Table 1.

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
−0.566	0.015	−0.146	0.032	0.142	−0.004	−0.008	−0.140	−0.144	0.772
0.182	−0.049	−0.662	−0.067	0.157	−0.012	−0.064	−0.207	0.667	0.069
0.248	−0.044	−0.610	−0.053	0.258	−0.005	0.018	0.231	−0.662	−0.060
−0.100	−0.011	−0.286	−0.360	−0.872	−0.015	0.045	−0.063	−0.106	0.017
−0.535	0.020	−0.172	0.123	0.114	−0.037	0.026	−0.547	−0.146	−0.577
0.165	−0.015	−0.140	0.906	−0.332	0.038	0.067	−0.074	−0.043	0.098
−0.507	0.013	−0.183	0.154	−0.039	0.034	−0.107	0.753	0.230	−0.227
0.001	−0.573	0.052	0.044	−0.036	−0.664	−0.473	−0.011	−0.039	0.008
−0.063	−0.582	0.013	−0.024	0.049	−0.075	0.799	0.072	0.071	−0.006
−0.018	−0.572	0.034	−0.018	−0.015	0.741	−0.339	−0.064	−0.040	−0.011

Alongside Figures 1–3, the extracted components provide insights into potential predictors of any attribute relating to the data in Table 1. Each of the four panels in Figure 4 exhibits clear cases of overlapping which, as noted earlier, potentially lead to data swamping, masking, model over-fitting or under-fitting. In the next sub-section, we demonstrated how these issues can be addressed via the SMA algorithm.

3.3. Supervised Modelling

For supervised modelling, we deploy ANN using standard back-propagation, allowing for flexibility in settings such as the threshold and the learning rate and influencing the values of $\hat{\mathcal{L}}_{tr,ts}$ and $\mathbb{E}[\Delta]$ in SMA Algorithm 1. As noted above, the target variable, CLS, was tweaked to give two instances—binary and trinomial. Equation (8) takes a non-linear form, with λ representing CLS. Its weights are initialised to small random values and updated via an optimisation algorithm in response to estimates of error on both the training and test datasets and all ten inputs were appropriately scaled. Two ANN models were trained: one on binary and the other on a trinomial class instance.

Each of the two models was trained and tested multiple times on random samples via the SMA Algorithm in search of optimal solutions. Two of the key parameters of an ANN model are the threshold and the learning rate. The former is a numeric value that specifies the stopping criteria based on the threshold for the partial derivatives of the error function. The smaller this value is, the deeper the model learns from the current training data, hence risking over-fitting the data. The latter describes the rate at which the weights are updated during the training process. Again, the gap between w_i and w_{i+1} has an impact on the output of the model. In the next sub-section, we explore the impact of these parameters on ANN modelling and we demonstrate the applicability of the SMA algorithm in model selection.

3.3.1. Thresholding and Learning Rate

Table 4 presents selected results from multiple ANN models based a range of thresholds and different training and testing samples. The decision on the value of the threshold to use in any particular application requires a good understanding of the data as well as the underlying domain knowledge. Usually, the classification of degree awards would be done by a specially delegated body, such as SILPA [12], which in this case will have influenced the outcomes of how the *CLS* variable is defined. One way to assess the impact of such decisions is to look at the predictive patterns based on the current segmentation. It is for that purpose that we deploy the SMA Algorithm.

To further illustrate foregoing assertions, we consider the training and testing errors in Table 4. Typically, we expect to see the relationship $\psi_{D,TEST} \gg \psi_{D,TRN}$ holding, i.e., a positive $\mathbb{E}[\Delta]$. Across the six samples in the table, the two panels in Figure 5 align with this position for the two-class case and diverge in three places for the three-class case.

Table 4. Selected model performances from competing models through the SMA algorithm.

Model ($\hat{\mathcal{L}}_{tr,ts}$)	Threshold	$\psi_{D,TRN}$	$\psi_{D,TEST}$	$\mathbb{E}[\Delta]$	Sample $[x_{\nu,\tau}]$	Sample $[x_{\bar{\nu},\tau}]$
ANN−Bin−1	0.50	0.02926	0.02764	−0.001618	$S_{tr} = 2529$	$S_{ts} = 615$
ANN−Tri−1	0.50	0.28143	0.31159	0.030157	$S_{tr} = 3006$	$S_{ts} = 138$
ANN−Bin−2	0.40	0.01979	0.03074	0.010950	$S_{tr} = 2526$	$S_{ts} = 618$
ANN−Tri−2	0.40	0.27945	0.29552	0.016063	$S_{tr} = 2809$	$S_{ts} = 335$
ANN−Bin−3	0.25	0.02228	0.02852	0.006242	$S_{tr} = 2513$	$S_{ts} = 631$
ANN−Tri−3	0.25	0.28689	0.25738	−0.029507	$S_{tr} = 2670$	$S_{ts} = 474$
ANN−Bin−4	0.10	0.00913	0.01757	0.008437	$S_{tr} = 2518$	$S_{ts} = 626$
ANN−Tri−4	0.10	0.28283	0.26737	−0.015464	$S_{tr} = 2482$	$S_{ts} = 662$
ANN−Bin−5	0.05	0.00434	0.00652	0.0021791	$S_{tr} = 2531$	$S_{ts} = 613$
ANN−Tri−5	0.05	0.27599	0.29562	0.019636	$S_{tr} = 2366$	$S_{ts} = 778$
ANN−Bin−6	0.01	0.00201	0.01801	0.016000	$S_{tr} = 2478$	$S_{ts} = 666$
ANN−Tri−6	0.01	0.28493	0.26580	−0.019129	$S_{tr} = 2211$	$S_{ts} = 933$

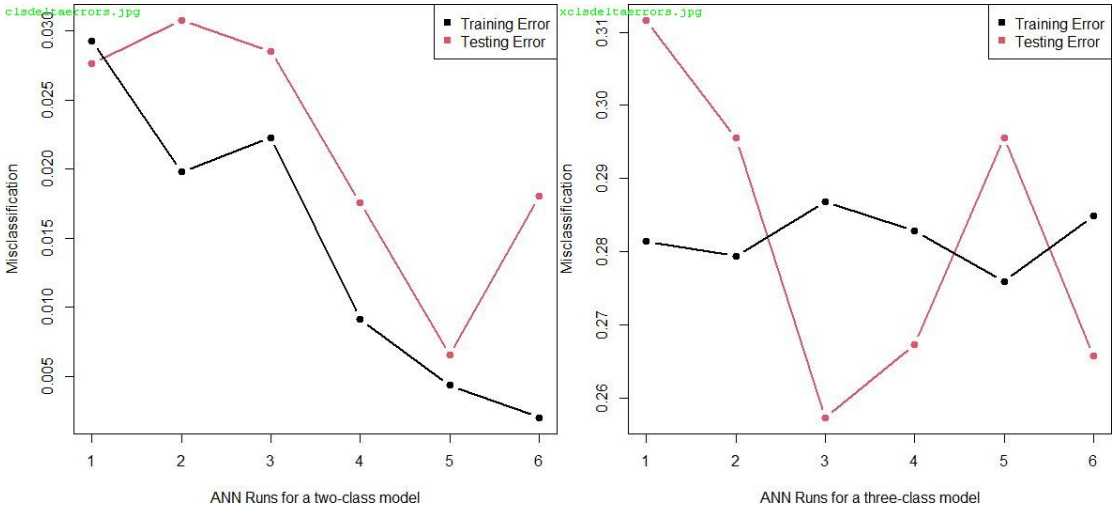

Figure 5. Training and testing error patterns for the two classes (LHS) and three classes (RHS).

The best results for the two-class case are given in Figure 6, showing the performance at both the hidden layer and output levels with the two classes clearly separated. This performance does not necessarily reflect the optimal architecture for the ANN, it is only that it presents a clear demarcation of the classes as defined. Significantly increasing the number of samples and assessing through the SMA Algorithm, more consistent patterns can be obtained.

Creating an additional GPA class, based on Equation (5), yielded the two panels in Figure 7. Like in the binary case above, they illustrate the performance of the testing model at the hidden layer and output node levels. The misclassification error for the test model in this case was 28%, which can be inferred from the separation in both panels. As noted above, the decision to discretise the variable *CLS* into three categories has a direct impact on the training and testing errors. The massive difference in accuracy between the two- and three-class cases underlines the importance of adopting interdisciplinary approaches for educationists and data scientists to work together.

In both cases—binomial and trinomial—the main concern, as is with all cases of predictive modelling, is data over-fitting or indeed, data under-fitting. Setting the parameters at steps 4, 8 and 10 as well as deciding on the sizes of S_{tr} and S_{ts} at steps 9 and 10 and setting the magnitude of $\mathbb{E}[\Delta]$ at step 20 all have a direct impact on model complexity. While there was no computational stress on this application, due to the relatively small data dimension, other applications with significantly higher numbers of observations and attributes are likely to demand more computational resources.

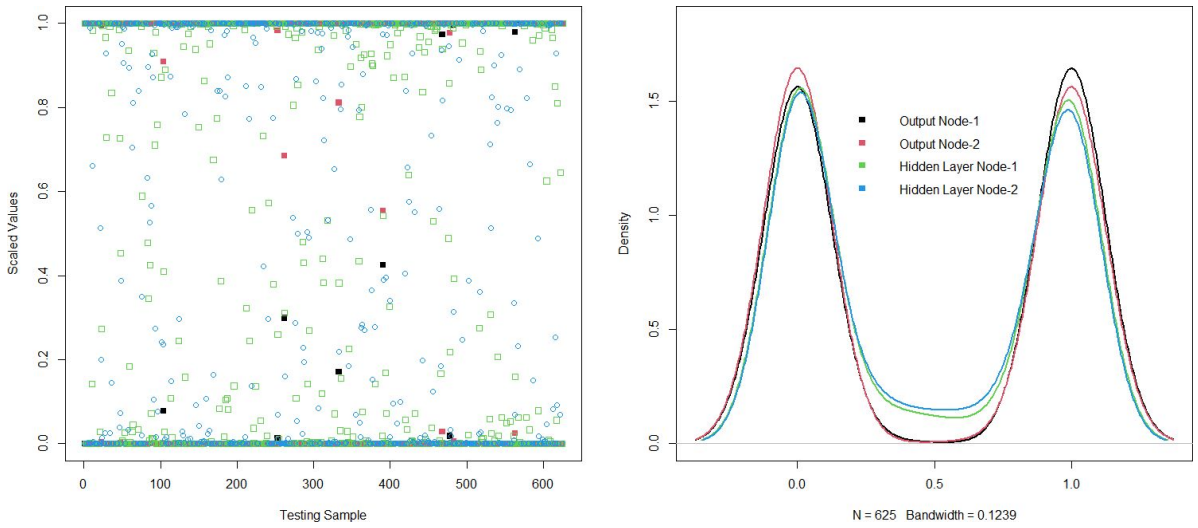

Figure 6. Neural network two-class test results.

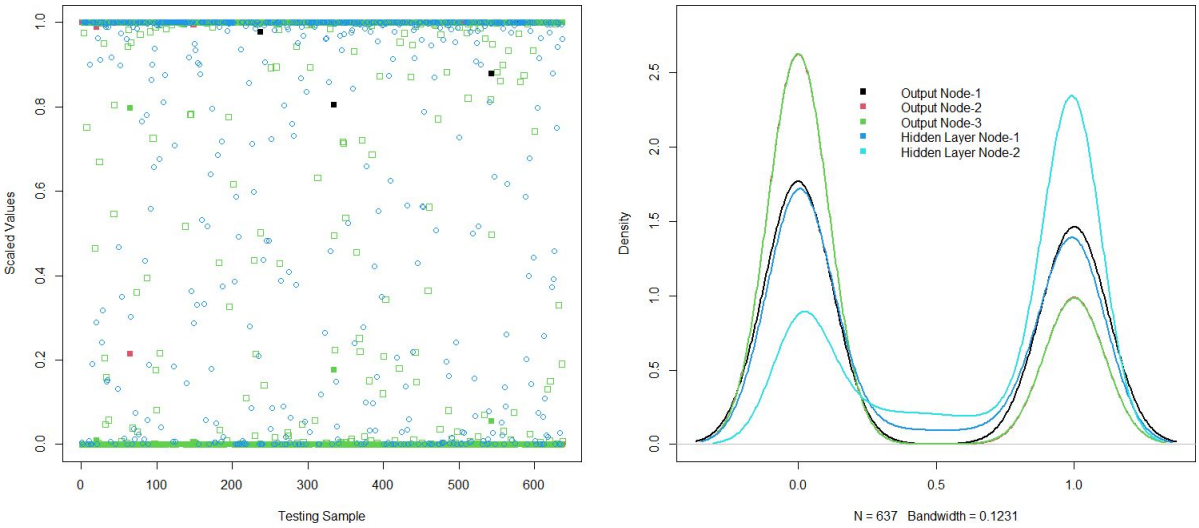

Figure 7. Neural network three-class test results.

4. Contribution to Knowledge and Discussion

This paper's novelty is two-fold, technical and application based. Model optimisation continues to be the focal point for researchers across applications. As stated in Section 2.2, model outcomes are conditional on the data used, the robustness of the statistical analysis and the interpretation of the results, as these combine to form a tool for knowledge extraction from data. On the other side, while good quality education is a well-acknowledged foundation of sustainable development, innovation and creativity, there is still no consensus around the world as to how such goals are to be attained. Hence, the search for what is the right breaking point for university pedagogy innovations continues to attract the attention of educationists and other researchers in both natural and social sciences. It is in this context that the novelty of this paper is attributed to an interdisciplinary approach—combining the pedagogy and data analytics. The results in Section 3 are presented to reflect the aims and objectives outlined in Section 1.2 and the gap challenges in Section 1.3. The paper's contribution to knowledge and general discussions are summarised below.

4.1. Contribution to Knowledge

The technical side of contribution to knowledge is embedded in the SMA algorithm's potential to address data randomness and concept drift. By minimising such variations and

yielding consistent results across samples, the algorithm provides enhanced steps towards the application of machine learning techniques in studying students' performance. The latter relates to identifying potential triggers of pedagogical performance, which aligns with SDG #4—Quality Education—and provides insights into the attainment levels of other SDGs in the UAE.

Findings from the paper are likely to draw the attention of CHEDS [9], the Standards for Institutional Licensure and Program Accreditation [12] and other education stakeholders in the UAE and elsewhere, to constantly monitor the dynamic nature of the relationships between input and output variables, through interdisciplinary engagements. Furthermore, while this application focused on knowledge gaps in higher education management and pedagogy, in the UAE, its approach is readily adaptable to applications outside the education sector and beyond the UAE.

4.2. Discussion

The SMA algorithm was presented in an adaptable form to handle both unsupervised and supervised modelling, in a setting that allows data randomness and concept drift to be captured, measured and assessed within the modelling process. The binomial and trinomial scenarios in Section 2.4.3 were selected via cross-validation involving a range of discretisation levels from 2 to 7, based on the assessment criterion at step #20 of the algorithm. Since the algorithm is data-dependent, we recommend that different applications experiment with as many adjustable parameters as possible. The influence of the initial parameters should be monitored via steps #5–#6 and step #20 for the final results.

Set to address the practical issues that practitioners encounter while making data-driven decisions and the intricacy of relationships among data attributes in the BD era [20,30], this paper focused on a real-life scenario of the challenges and opportunities we face in what is an increasingly interdisciplinary and globalised environment. This paper sought to highlight paths towards a meaningful application of data analytics techniques from simple to complex settings. The intricate relationships among attributes describe the invariant conditions that practitioners in the two overlapping fields of data science and education must identify, as they epitomise the meaningfulness of the concept of learning rules from data. The disparate performance accuracies in the two models, exhibited in Table 4 and the related graphical illustrations, reflect the well-documented issue of concept drift [10,11]—an expression referring to changes in the statistical properties of the target variable, such as CLS, over time. It is expected that the findings of this paper will inspire interdisciplinary engagements and provide good input to education stakeholders both within the UAE and beyond. This work is also expected to stimulate further discussions on a wide range of topics, such as the impact of internships on performance and employability, performance differentiation-based specialisation or mode of study.

Students' performance continues to attract the interests of many researchers, both within and outside the education sector. In the education sector, the main focus has been on student retention and satisfaction [31], on the one side, and employability, i.e., building bridges with the industry on the other [32]. Industrialists are keen to establish the type of apprenticeships that best suit their core businesses [33,34]. Between the two sectors, academic engagement and the ultimate impact of the relationship between the academia and the labour market has been central. The two sectors are both vulnerable to concept drift—a direct or indirect consequence of changing definitions or data properties over time, causing predictions based on previous or current definitions to become less accurate over time. It is imperative that educationists identify performance triggers among students, a knowledge that impinges not only on the students' employability, but also on their potential productivity. For example, apprenticeships can be directly linked to internships in that students on placement provide managers with exemplars that they may want to sustain [33]. Innovative undergraduate programs in Data Science have the potential to expand the horizon of learners in many ways.

This work was motivated by recent developments in data generation, processing and their impact on tackling societal challenges, particularly in the education sector. Its aims and objectives in Section 1.2 were inspired by the scientific quest to address issues of data randomness and concept drift in education. The applications in Section 3 focused on data visualisation, unsupervised modelling and supervised modelling. The ultimate purpose of these applications was to uncover hidden information in education data and potentially utilise this knowledge in enhancing decision making in the sector. Data visualisation exhibited patterns that tell stories about various aspects of the data in Table 1—quite useful information for end users and decision makers, who typically prefer interpretable and understandable solutions.

The title of the paper reflects the complexity of making inference based on datasets. As it is conventional in statistical hypothesis testing, where we construct and verify hypotheses by investigating available data before drawing inferences, unsupervised and supervised modelling relied on the absence or presence of prior information on the underlying structures in the data in Table 1. Without the tweaked variable *CLS*, we applied the SMA algorithm to identify naturally arising components in the students' data. This kind of natural grouping provides insights into the overall patterns across the sector. Again, the interpretability and understandability of these patterns to end users and decision makers are crucial. Labelling the historical data in Table 1 made it possible to use the SMA algorithm with a predictive modelling technique, such as the ANN, to make performance predictions. Again, these predictions can be quite useful for end users and decision makers, as they provide a benchmark for evaluating the education sector, not only by looking at how they compare with actual performance, but also by guiding interim decisions and policies.

It is also imperative to highlight this study's limitations and some of its potential issues. The study used 3144 observations gathered over the time period 2005–2016. CHEDS [9] statistics show that the UAE higher education sector has over 100 higher and further education providers, enrolling approximately 140,000 students. The federal government institutions—United Arab Emirates University (UAEU), Zayed University and the Higher Colleges of Technology—account for approximately 30% of total student enrolment (approximately 43,000 students). In comparison to the sample used in the study, it is clear that given the dynamics in the sector, regular pedagogical analyses need to be carried out. Most importantly, COVID-19 appears to have had a lasting impact on the way education is delivered, which might invalidate some of the findings of this study. However, going forward, data randomness and concept drift will continue to hinge on model optimisation—a classical challenge in making data-driven decisions [7]. Hence, enhancements of the main ideas of this paper derive from the premises that the foregoing relationships constitute a potential source of Big Data (BD) flowing and shared across sectors. Finally, as the Big Data era promotes more trust in data and algorithms than human judgment, we are called upon to address ethical and privacy issues, which this study has not addressed.

5. Concluding Remarks

This paper sought to address the problem space in Section 1.2 via the four objectives. An ANN model was applied to learn the rules for mapping the numerical (credit and GPA) inputs to two types of output variables—in two–three-class cases. The SMA Algorithm was applied to improve the learning process, i.e., to avoid exploding gradients. This paper set off from the premises that since education generates large volumes of data, its interpretations and proper utilisation have a direct impact on societal innovation and productivity. The first three objectives were clearly met and it is expected that the fourth objective, set to motivate unified initiatives of educationists and data scientists towards interdisciplinary applications of this nature and others, will be fulfilled in the course of time.

Whilst our interests are increasingly drawn to the unification of concepts in computing and statistics, our findings reveal natural links among different majors in the students' data. These findings should be pooled back into curriculum designs in order to provide

students with the necessary skills to comprehend their learning habitat, purpose and outcome. It is expected that this work will lead to comparative studies across universities in the Emirates, providing an opportunity for data sharing. It is expected that the paper's findings will inspire interdisciplinary engagements and provide good input to education stakeholders, including the relevant authorities in the UAE, to constantly monitor the dynamic nature of the relationships between input and output variables, through interdisciplinary engagements.

In the modern era of Big Data, there is scope for extending this study into a range of potential future directions. For example, ideas from metaheuristic optimisation techniques such as particle swarm optimisation [35,36] could be adapted to capture students' performances on shorter and more regular periodic intervals. Apparently, capturing progress in this way would require curricula adaptation, possibly with universities or departments setting targets and treating every student's performance as a "particle", moving at predetermined velocity during studies.

Author Contributions: As a result of previous joint work, both authors equally contributed to this work. K.S.M. carried out most of the data cleaning and automated selection and R.A.S. provided the raw data and many of the insights into designing the analyses layout based on their experiences with the education system in the UAE. Problem conceptualisation was initiated by R.A.S. and K.S.M. proposed the methodology. Original data compilation was performed by R.A.S. followed by a joint data preparation process. Writing up was shared across several iterations of reviews and editing. On average, both authors equally contributed to the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was not supported by any grant, but it is an outcome of ordinary Research and Scholarly Activities (RSA) allocation to each of the two authors by their respective institutions.

Institutional Review Board Statement: Ethical review and approval was not applicable for this study as it did not involve humans or animals.

Informed Consent Statement: Informed consent was not applicable for this study as it did not involve humans.

Data Availability Statement: As noted in Section 2.1, the data attributes used in this study were obtained via a semi-automated random selection and cleaning process by the authors. They were reformatted to fit in with the adopted modelling strategy, hence, the data are only available from the authors, who retained both the raw and modified copies, should they be requested.

Acknowledgments: This paper is part of ongoing initiatives towards the Big Data modelling of the Sustainable Development Goals (BDMSDG) and the application of the DSF both authors have been involved in over the last three years. We would like to thank many individuals and institutions who have discussed these initiatives with us at different stages of development. We particularly acknowledge the role of the Data Intensive Research Initiative of South Africa (DIRISA), through the South African Council for Scientific and Industrial Research (CSIR), who have invited us a couple of times to Pretoria to present our findings. We also acknowledge the presentation opportunity we have had with the Joint Support-Center for Data Science Research (DS), through the Japanese Polar Environment Data Science Center (PEDSC), the United Nations World Data Forum (UNWDF) and the Sussex Sustainability Research Programme (SDG interactions) of the University of Sussex. Most importantly, we are grateful to CHEDS in the MoE of the UAE for providing the raw data for this work.

Conflicts of Interest: Both authors declare that there are no competing interests in publishing this paper, be they financial or non financial.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Networks
BD	Big Data
BDMSDG	Big Data Modelling of Sustainable Development Goals
CAA	Commission for Academic Accreditation
CSIR	Council for Scientific and Industrial Research
DIRISA	Data Intensive Research Initiative of South Africa
DSF	Development Science Framework
DV	Data Visualisation
EDA	Exploratory Data Analysis
GPA	Grade Point Average
MoE	Ministry of Education
PCA	Principal Component
PEDSC	Polar Environment Data Science Centre
SDG	Sustainable Development Goals
SILPA	Standards for Institutional Licensure and Program Accreditation
SMA	Sample–Measure–Assess
UAE	United Arab Emirates
UNWDF	United Nations World Data Forum

References

1. Costa, E.B.; Fonseca, B.; Santana, M.A.; de Araújo, F.F.; Rego, J. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Comput. Hum. Behav.* **2017**, *73*, 247–256.
2. Wilson, K. What does it mean to do teaching? A qualitative study of resistance to Flipped Learning in a higher education context. *Teach. High. Educ.* **2020**, 1–14, doi:10.1080/13562517.2020.1822312.
3. Hua Leong, F.; Marshall, L. Modeling engagement of programming students using unsupervised machine learning technique. *GSTF J. Comput. (JoC)* **2018**, *6*, 1-6.
4. Brooks, C.; Erickson, G.; Greer, J.; Gutwin, C. Modelling and quantifying the behaviours of students in lecture capture environments. *Comput. Educ.* **2014**, *75*, 282–292.
5. Miguéis, V.L.; Freitas, A.; Garcia, P.J.V.; Silva, A. Early segmentation of students according to their academic performance: A predictive modelling approach. *Decis. Support Syst.* **2018**, *115*, 36–51,
6. Domínguez Figaredo, D. Data-Driven Educational Algorithms Pedagogical Framing. *Revista Iberoamericana de Educación a Distancia* **2020**, *23*, 65–84.
7. Mwitondi, K.S.; Said, R.A. A data-based method for harmonising heterogeneous data modelling techniques across data mining applications. *J. Stat. Appl. Probab.* **2013**, 2, 293–305.
8. Zenisek, J.; Holzinger, F.; Affenzeller, M. Machine learning based concept drift detection for predictive maintenance. *Comput. Ind. Eng.* **2019**, *137*, 106031.
9. CHEDS. *Center For Higher Education Data and Statistics*; Ministry of Education: Dubai, UAE, 2018.
10. Žliobaitė, I.; Pechenizkiy, M.; Gama, J. An Overview of Concept Drift Applications. In *Big Data Analysis: New Algorithms for a New Society*; Japkowicz, N., Stefanowski, J., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 91–114.
11. Tsymbal, A.; Pechenizkiy, M.; Cunningham, P.; Puuronen, S. Dynamic integration of classifiers for handling concept drift. *Inf. Fusion* **2008**, *9*, 56–68.
12. SILPA. *Standards for Institutional Licensure and Program Accreditation*; Ministry of Education: Dubai, UAE, 2019.
13. Mwitondi, K.S.; Moustafa, R.E.; Hadi, A.S. A Data-Driven Method for Selecting Optimal Models Based on Graphical Visualisation of Differences in Sequentially Fitted ROC Model Parameters. *Data Sci. J.* **2013**, *12*, WDS247–WDS253.
14. Saggi, M.K.; Jain, S. A survey towards an integration of big data analytics to big insights for value-creation. *Inf. Process. Manag.* **2018**, *54*, 758–790.
15. Reyes, J.A. The skinny on big data in education: Learning analytics simplified. *Techtrends Tech Trends Springer* **2015**, *59*, 75–80.
16. Zhou, L.; Pan, S.; Wang, J.; Vasilakos, A.V. Machine learning on big data: Opportunities and challenges. *Neurocomputing* **2017**, *237*, 350–361.
17. Xu, Y.; Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J. Anal. Test.* **2018**, *2*, 249–262.
18. Chen, S.; Dorn, S.; Lell, M.; Kachelrieß, M.; Maier, A. *Manifold Learning-Based Data Sampling for Model Training*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 269–274.
19. Mwitondi, K.; Munyakazi, I.; Gatsheni, B. A robust machine learning approach to SDG data segmentation. *J. Big Data* **2020**, *7*, doi:10.1186/s40537-020-00373-y.

20. Mwitondi, K.; Munyakazi, I.; Gatsheni, B. Amenability of the United Nations Sustainable Development Goals to Big Data Modelling. In Proceedings of the International Workshop on Data Science-Present and Future of Open Data and Open Science, Joint Support Centre for Data Science Research, Mishima Citizens Cultural Hall, Mishima, Shizuoka, Japan, 12–15 November 2018.
21. Mwitondi, K.; Munyakazi, I.; Gatsheni, B. An Interdisciplinary Data-Driven Framework for Development Science. In Proceedings of the DIRISA National Research Data Workshop, CSIR ICC, Pretoria, South Africa, 19–21 June 2018.
22. Drori, I.; Krishnamurthy, Y.; Lourenco, R.; Rampin, R.; Cho, K.; Silva, C.; Freire, J. Automatic Machine Learning by Pipeline Synthesis using Model-Based Reinforcement Learning and a Grammar. *arXiv* **2019**, arXiv:cs.LG/1905.10345.
23. Bo, L.; Wang, L.; Jiao, L. Feature Scaling for Kernel Fisher Discriminant Analysis Using Leave-One-Out Cross Validation. *Neural Comput.* **2006**, *18*, 961–978.
24. Galkin, F.; Aliper, A.; Putin, E.; Kuznetsov, I.; Gladyshev, V.N.; Zhavoronkov, A. Human microbiome aging clocks based on deep learning and tandem of permutation feature importance and accumulated local effects. *bioRxiv* **2018**, doi:10.1101/507780.
25. Mwitondi, K.S.; Said, R.A.; Zargari, S.A. A robust domain partitioning intrusion detection method. *J. Inf. Secur. Appl.* **2019**, *48*, 102360.
26. Looney, C.G. *Pattern Recognition Using Neural Networks: Theory and Algorithms for Engineers and Scientists*; Oxford University Press: New York, NY, USA, 1997.
27. Webb, A. *Statistical Pattern Recognition*; Wiley: London, UK, 2005.
28. Lawrence, A.J. Deletion Influence and Masking in Regression. *J. R. Stat. Society. Ser. B (Methodol.)* **1995**, *57*, 181–189.
29. Bendre, S.M. Masking and swamping effects on tests for multiple outliers in normal sample. *Commun. Stat. Theory Methods* **1989**, *18*, 697–710.
30. Parsons, M.A.; Øystein Godøy.; LeDrew, E.; de Bruin, T.F.; Danis, B.; Tomlinson, S.; Carlson, D. A conceptual framework for managing very diverse data for complex, interdisciplinary science. *J. Inf. Sci.* **2011**, *37*, 555–569, doi:10.1177/0165551511412705.
31. Johnson, S.R.; Stage, F.K. Academic Engagement and Student Success: Do High-Impact Practices Mean Higher Graduation Rates? *J. High. Educ.* **2018**, *89*, 753–781, doi:10.1080/00221546.2018.1441107.
32. Rienties, B.; Toetenel, L. The impact of learning design on student behaviour, satisfaction and performance: A cross-institutional comparison across 151 modules. *Comput. Hum. Behav.* **2016**, *60*, 333–341,
33. Lerman, R. Do firms benefit from apprenticeship investments? *IZA World Labor* **2019**, doi:10.1371/journal.pone.0192976.
34. Di Meglio, G.; Barge-Gil, A.; Camiña, E.; Moreno, L. Knocking on Employment´s Door: Internships and Job Attainment. *Munich Personal RePEc Archive* **2019**. Available online: https://mpra.ub.uni-muenchen.de/95712/1/MPRA_paper_95712.pdf (accessed on 15 July 2021).
35. Kennedy, J.; Eberhart, R. Particle swarm optimization. In Proceedings of the ICNN'95—International Conference on Neural Networks, Perth, WA, Australia, 27 November–1 December 1995; Volume 4, pp. 1942–1948
36. Shi, Y.; Eberhart, R. A modified particle swarm optimizer. In Proceedings of the 1998 IEEE International Conference on Evolutionary Computation Proceedings, IEEE World Congress on Computational Intelligence (Cat. No.98TH8360), Anchorage, AK, USA, 4–9 May 1998; pp. 69–73.